# 데이터 분포 통계를 이용한 CSV 형식의 공공데이터 도메인 판별 모델에 관한 연구

정하나[O], 김재웅[*], 이윤열[**], 채의근[**], 정영석[***]
[O]공주대학교 대학원 컴퓨터공학과,
[*]공주대학교 소프트웨어학과,
[**]공주대학교 컴퓨터공학과,
[***]공주대학교 대학원 컴퓨터공학과
e-mail: konghanaj@gmail.com[O], jwkim@kongju.ac.kr[*],
{alphaone, ygchae}@kongju.ac.kr[**], merope@kongju.ac.kr[***]

# A Study on Domain Discrimination Model for CSV Format Public Data Using Data Distribution Statistics

Ha-Na Jeong[O], Jae-Woong Kim[*], Yun-Yeol Lee[**], Yi-Geun Chae[**], Young-Suk Chung[***]
[O]Dept. of Computer Engineering, Kongju National University,
[*]Dept. of Software, Kongju National University,
[**]Dept. of Computer Engineering, Kongju National University,
[***]Dept. of Computer Engineering, Kongju National University

● 요 약 ●

정부는 공공데이터의 품질 관리를 위하여 공공데이터 품질관리 수준평가를 진행하여 공공데이터 품질을 관리하고 있다. 파일 형식의 공공데이터를 진단 시 품질진단 담당자가 대량의 파일데이터를 필드명과 필드 내 데이터에 의존하여 수작업으로 도메인을 판단하여 진단한다. 때문에 품질진단의 정확성을 신뢰하기 어렵고 진단에 많은 시간이 소요된다. 본 논문은 파일형식의 공공데이터 품질진단의 정확성을 확보하고 진단 소요 시간을 단축하기 위해 데이터 분포 통계를 이용한 CSV 형식의 공공데이터 도메인 판별 모델을 제안하였다. 제안된 모델을 적용하면 공공데이터 품질의 정확성을 향상하고 진단 소비 시간을 단축시킬 것으로 기대된다.

**키워드:** 공공데이터(Public data), 데이터품질(Data quality), 품질개선(Quality improvement), 데이터 품질진단(Data quality diagnosis), 데이터베이스(Database)

## I. Introduction

정부가 보유하고 있는 공공데이터를 민간에 개방하는 오픈데이터 정책은 전 세계적으로 화두에 오르고 있다. 이에 대한민국 정부는 국민경제 활성화에 기여의 목적으로 2013년 공공데이터의 제공 및 이용 활성화에 관한 법률을 제정하였다[1]. 정부는 2023년 4월 현재 기준 공공데이터 포털(data.go.kr)을 통해 오픈API 10,906건, 파일데이터 58,118건, 표준데이터셋 9,337건을 개방하고 있다[2]. 그러나 꾸준히 확대되고 있는 공공데이터 개방에도 불구하고 공공데이터의 활용도는 기대에 미치지 못하고 있다. 활용도 저하의 주요 원인으로는 공공데이터 품질 관리의 미흡이 제기되고 있다[3]. 정부는 이를 해결하고자 한국지능정보화진흥원을 통해 공공데이터 품질관리 수준평가를 실시하여 공공 개방데이터의 오류에 대한 진단 및 개선 작업을 수행하고 있다. 그러나 데이터 필드명과 필드 내 데이터에 의존하여 수작업으로 데이터의 도메인을 판별해야 하는 특성을 가지는 공공 개방데이터의 오류진단은 비용과 시간을 많이 소비한다. 또한 품질진단의 정확성을 보장할 수 없다. 이를 해결하기 위해 본 논문에서는 데이터 분포 통계를 이용한 CSV(Comma-Separated Values) 형식 공공데이터의 도메인 판별 모델에 관한 연구를 진행하였다.

## II. Preliminaries

### 1. Related works

#### 1.1 공공데이터 제공 형식

공공데이터법 제2조 제3호에 따르면 공공데이터는 기계판독이 가능한 형태로 제공해야 한다[4]. 이는 한컴오피스, Micro Office 등의 상용 소프트웨어를 통해 수정/변환/추출 등 가공할 수 있는 형태의 데이터를 뜻한다. 따라서 공공데이터는 CSV, JSON, XML, XLS 등의 오픈포맷 형태의 데이터여야만 한다. 정부는 공공데이터 포털을 통해 여러 오픈포맷 형태의 데이터를 개방하고 있다. 아래 Table 1. 은 공공데이터 포털에서 공공데이터 개방 시 사용되고 있는 주요 데이터 포맷별 공공데이터 개수이다.

Table 1. Public Open Data Major File Format

| File Format | Amount |
|---|---|
| CSV | 44,360 |
| JSON | 1,760 |
| XML | 4,800 |
| HWP | 2,036 |
| XLSX | 1,468 |
| XLS | 1,168 |

현재 기준 공공데이터 포털에 개방되어있는 데이터는 총 78,361건이다. 이 중 CSV 포맷 데이터가 44,360건으로 약 56.6%를 차지한다.

## III. The Proposed Scheme

본 논문에서는 공공데이터의 도메인별 데이터 분포 평균치를 산출하여 도메인 판별에 이용하는 모델을 제안한다. 아래의 Fig. 1. 은 본 논문에서 제안하는 데이터 분포 통계를 이용한 CSV 형식 공공데이터의 도메인 판별 모델의 프로세스이다.

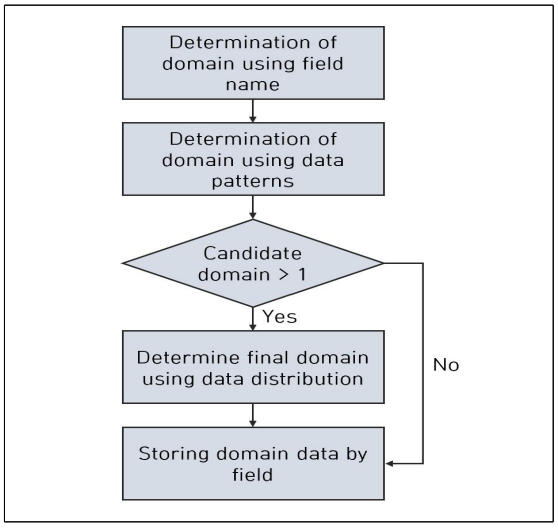


Fig. 1. Process of Domain Discrimination Model for CSV Format Public Data Using Data Distribution Statistics

첫 번째, CSV 파일의 필드명을 이용하여 도메인을 1차로 판별한다.

두 번째, 판별된 후보 도메인을 필드 내 데이터 패턴 정보를 이용하여 2차로 검토 및 판별한다.

세 번째, 만약 2차 판별 후 후보 도메인이 1개 이하라면 해당 필드에 대한 도메인 데이터를 저장한다.

네 번째, 만약 2차 판별 후 후보 도메인이 1개 초과라면 필드 내 데이터 분포와 공공데이터 도메인별 평균 데이터 분포를 비교하여 도메인을 판별한다. 공공데이터의 도메인별 평균 데이터 분포는 분류체계별로 임의의 CSV 포맷 데이터를 선정하여 데이터 내 각 필드의 도메인을 수동으로 분류 후 도메인별 평균 데이터 분포를 분석하여 산출한다. 이후 판별된 1개의 도메인을 해당 필드에 대한 도메인 데이터로 저장한다.

위의 프로세스를 통하여 CSV 형식 공공데이터의 각 필드별 도메인을 판별하여 저장한다. 이후 필드별 도메인 데이터를 공공데이터 품질진단에 이용한다.

## IV. Conclusions

본 논문은 CSV 형식 공공데이터의 도메인별 평균 데이터 분포를 산출하여 CSV 형식 공공데이터의 도메인을 판별하는 모델을 제안하였다. 제안한 논문을 활용하면 공공 파일데이터 품질진단 시 진단 결과의 정확도의 향상과 진단 소비 시간 단축이 가능할 것으로 기대된다. 향후 CSV 형식의 공공데이터를 활용하여 도메인별 평균 데이터 분포를 산출하고 제안된 모델을 구현하여 진단 결과의 정확성에 관한 연구를 진행할 예정이다.

## REFERENCES

[1] Ywhong, "A study on the invigorating strategies for open government data", The Korean Data and Information Science Society, Vol 25, No. 4, pp. 769-777, Aug. 2014. DOI: https://doi.org/10.7465/jkdi.2014.25.4.769

[2] Ministry of the Interior and safety, Public data portal, *https://www.data.go.kr*.

[3] ShPark, khLee, ayLee. "An Empirical Study on the Effects of Source Data Quality on the Usefulness and Utilization of Big Data Analytics Results" Korea Data Strategy Society, Vol 24, No 4, pp. 197-214, 2017. DOI: doi.org/10.21219/jitam.2017.24.4.197

[4] Ministry of Government Legislation, Act on Promotion of Provision and Use of Public Data, *https://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률*.